AFRL-SN-WP-TR-2001-1106

FEATURE EXTRACTION USING AN INFORMATION THEORETIC FRAMEWORK

JOSE C. PRINCIPE

COMPUTER NEUROENGINEERING LABORATORY
EB 451, BLDG #33
UNIVERSITY OF FLORIDA
GAINESVILLE, FL 32611

DECEMBER 1999

FINAL REPORT FOR PERIOD OF 25 FEBRUARY 1997 – 01 OCTOBER 1999

20011218 119

SENSORS DIRECTORATE
AIR FORCE RESEARCH LABORATORY
AIR FORCE MATERIEL COMMAND
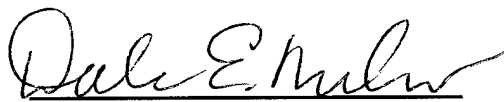WRIGHT-PATTERSON AIR FORCE BASE, OH 45433-7318

NOTICE

USING GOVERNMENT DRAWINGS, SPECIFICATIONS, OR OTHER DATA INCLUDED IN THIS DOCUMENT FOR ANY PURPOSE OTHER THAN GOVERNMENT PROCUREMENT DOES NOT IN ANY WAY OBLIGATE THE US GOVERNMENT. THE FACT THAT THE GOVERNMENT FORMULATED OR SUPPLIED THE DRAWINGS, SPECIFICATIONS, OR OTHER DATA DOES NOT LICENSE THE HOLDER OR ANY OTHER PERSON OR CORPORATION; OR CONVEY ANY RIGHTS OR PERMISSION TO MANUFACTURE, USE, OR SELL ANY PATENTED INVENTION THAT MAY RELATE TO THEM.

THIS REPORT IS RELEASABLE TO THE NATIONAL TECHNICAL INFORMATION SERVICE (NTIS). AT NTIS, IT WILL BE AVAILABLE TO THE GENERAL PUBLIC, INCLUDING FOREIGN NATIONS.
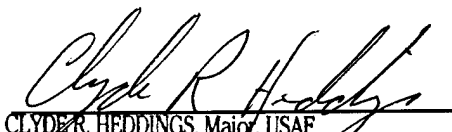
THIS TECHNICAL REPORT HAS BEEN REVIEWED AND IS APPROVED FOR PUBLICATION.

LOUIS A. TAMBURINO
Target Recognition Branch
Project Engineer

DALE E. NELSON, CHIEF
Target Recognition Branch

CLYDE R. HEDDINGS, Major, USAF
Deputy, Sensor ATR Technology Division
Sensors Directorate

Do not return copies of this report unless contractual obligations or notice on a specific document require its return.

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 074-0188

Public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing this collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden to Washington Headquarters Services, Directorate for Information Operations and Reports, 1215 Jefferson Davis Highway, Suite 1204, Arlington, VA 22202-4302, and to the Office of Management and Budget, Paperwork Reduction Project (0704-0188), Washington, DC 20503

1. AGENCY USE ONLY (Leave blank)	2. REPORT DATE December 1999	3. REPORT TYPE AND DATES COVERED Final Report, 02/25/1997 – 10/01/1999

4. TITLE AND SUBTITLE
Feature Extraction Using an Information Theoretic Framework

6. AUTHOR(S)
Jose C. Principe

5. FUNDING NUMBERS
C: F33615-97-1-1019
PE: 62301E
PR: ARPA
TA: AA
WU: 1U

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

COMPUTER NEUROENGINEERING LABORATORY
EB 451, BLDG #33
UNIVERSITY OF FLORIDA
GAINESVILLE, FL 32611

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)

SENSORS DIRECTORATE
AIR FORCE RESEARCH LABORATORY
AIR FORCE MATERIEL COMMAND
WRIGHT-PATTERSON AIR FORCE BASE, OH 45433-7318
POC: Louis A. Tamburino, AFRL/SNAT, 937-255-1115 x4389

10. SPONSORING / MONITORING AGENCY REPORT NUMBER
AFRL-SN-WP-TR-2001-1106

11. SUPPLEMENTARY NOTES

12a. DISTRIBUTION / AVAILABILITY STATEMENT
Approved for public release; distribution unlimited.

12b. DISTRIBUTION CODE

13. ABSTRACT *(Maximum 200 Words)*
This report addresses the rejection of confusers, the last piece of the work conducted under the contract F33615-97-1-1019. The performance of the information theoretic feature extraction is elucidated and compared with the traditional (perceptrons and template matchers) classifiers in the Moving and Stationary Target Acquisition and Recognition (MSTAR) database. But no performance evaluation would be complete without assessing the quality of the new classifier in rejection to confusers. Therefore, the MSTAR database and the previous classifiers were utilized as the basis of comparison. Results are that the information theoretic feature extraction works at the same performance level as the very sophisticated support vector machine (SVM) for both misclassification error and rejection to confusers. The method should be more widely applied since its use transcends classification: it is a general method to create features that preserve as much information as possible with respect to a given response. It has also been shown that the same principle can be applied to classification and pose estimation, which shows the wide applicability of the technique.

14. SUBJECT TERMS

Pose estimator, SAR, Automatic Target Recognition, Information Theory, Feature Extraction

15. NUMBER OF PAGES
44

16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT Unclassified	18. SECURITY CLASSIFICATION OF THIS PAGE Unclassified	19. SECURITY CLASSIFICATION OF ABSTRACT Unclassified	20. LIMITATION OF ABSTRACT SAR

NSN 7540-01-280-5500

Standard Form 298 (Rev. 2-89)
Prescribed by ANSI Std. Z39-18
298-102

i

I. Executive Summary

This report addresses the rejection to confusers, the last piece of the work conducted under the contract F33615-97-1019. Recall that in the previous report we ellucidated the performance of the information theoretic feature extraction, and compared our method with the traditional (perceptrons and template matchers) classifiers in the MSTAR database. But no performance evaluation would be complete without assessing the quality of the new classifier in rejection to confusers. Therefore we utilized the MSTAR database and the previous classifiers as the basis of our comparison. We are happy to report that the information theoretic feature extraction (ITL) works at the same performance level as the very sophisticated support vector machine (SVM) for both misclassification error and rejection to confusers. However, we expect that our method will be more widely applied since its use transcends classification: it is a general method to create features that preserve as much information as possible w.r.t. to a given response. We have shown that the same principle can be applied to classification and pose estimation, which shows the wide applicability of the technique.

In order to make the report self contained we present an overview of the method and the results. I also include in the next page the list of paper published funded by this work.

Book Chapters

Principe, J., Xu D., Fisher J., Information Theoretic Learning, in <u>Unsupervised Adaptive Filtering</u>, Simon Haykin Editor, 265-319, Wiley, 2000.

Candocia F., Principe J., Superresolution with Learned Local Kernels, in <u>Multimedia Signal Processing</u>, Editors Larsen, Guan, Kung, CRC Press, 2000.

Papers in refereed Jornals

1999

Zhao Q., Principe J., Brennan V., Xu D., Wang Z., "Synthetic aperture radar automatic target recognition with three strategies of learning and representation", accepted to the special issue on ATR, <u>Optical Engineering</u>.

Zhao Q., Principe J., "Forming large margins with support vector machines for synthetic aperture radar automatic target recognition, submitted to <u>Optical Engineering</u>

Principe J., Xu D., Zhao Q., Fisher J. "Learning from examples with information theoretic criteria", accepted in <u>VLSI Signal Processing Systems</u>.

1998

Candocia F., Principe J., "Super-resolution of images based on local correlations", <u>IEEE Trans. Neural Networks,</u> vol 10, #2, 372-380.

Fancourt C., Principe J., "Competitive principal component analysis for locally stationary time series", in <u>IEEE Trans. Signal Proc.</u>, vol 46, #11, 3068-3082.

Principe J., Kim M., Fisher J., "Target detection in synthetic aperture radar (SAR) using artificial neural networks", <u>IEEE Trans. Image Proc.</u> special issue on neural networks, vol 7, #8, 1136-1149.

Conference Publications

1999

Brennan V., Principe J., "Multiresolution using Principal Component Analysis", submitted to ICASSP 2000.

Zhao Q., and Principe J., "Forming Large Margins with Support Vector Machines For Synthetic Aperture Radar Automatic Target Recognition", Automatic Target Recognition IX, vol 3718, 101-107, SPIE Conference, Orlando.

Xu D., Principe J.,"Training multilayer perceptrons layer by layer", Int. J. Conf. Neural Nets. (IJCNN), Washington.

Xu D. and Principe J., "Training multilayer perceptron layer by layer with information potential", Proc. IEEE Int. Conf. Acoustic Speech Signal Proc. ICASSP'99, paper 2465.

Principe J., Xu D., "Information Theoretic Learning using Renyi's quadratic entropy", in Proc. ICA'99, 407-412, Aussois, France.

1998

Fisher J., Principe J., "Blind source separation by interactions of output signals", IEEE Workshop on Sig. Proc.DSP98, Utah.

Zhao Q., Xu D., Principe J., "Pose estimation for SAR Automatic Target Recognition", Image Underst. Workshop, vol II, 827-831, Pacific Grove, Ca.

Boillot M., Principe J., "An algorithm development and support database for MSTAR", Image Underst. Workshop, vol II, 821-825, Pacific Grove, Ca.

Principe J., Zhao Q., Xu D., "A novel classifier architecture exploiting pose information for SAR/ATR", Image Underst. Workshop, vol II, 833-837, Pacific Grove, Ca.

Xu D., Principe J., "Learning from examples with mutual information", Proc. IEEE Workshop on NNSP, 155-164.

Xu D., Principe J., Fisher J., Wu H., A novel measure for independent component analysis, in Proc. ICASSP98, vol II 1161-1164.

Fisher J., Principe J., "A methodology for information theoretic feature extraction", in Proc. IJCNN 98 vol III 1712-1716.

Xu D, Fisher J., Principe J., "Mutual information approach to pose estimation", accepted in SPIE 98, Algorithms for synthetic aperture radar imagery V, Ed. Ed Zelnio, vol 3370, 218-229.

Candocia F., Principe J., "A method using multiple models to super-resolve SAR imagery", accepted in SPIE 98, Algorithms for synthetic aperture radar imagery V, Ed. Ed Zelnio, vol 3370, 197-207.

1997

Fancourt C., Principe J., "Soft competition principal component analysis using the mixture of experts", Proc. of the Image Understanding Workshop, 1071-1076, New Orleans.

Fisher J., Principe J., "A nonparametric methodology for information theoretic feature extraction", Proc. of the Image Understanding Workshop, 1077- 1084, New Orleans.

Candocia F., Principe J., "A self-organizing principle for segmenting and super-resolving ISAR images", Proc. of the Image Understanding Workshop, 1161-1166, New Orleans.

Yen L., Principe J., "Target detection in UWB images using temporal fusion", Proc. of the Image Understanding Workshop, vol II 1155-1160, New Orleans.

Fisher J., Principe J., "Entropy manipulation of arbitrary nonlinear mappings", in Proc. IEEE Workshop NNSP7, 14-23, Amelia Island.

I. Introduction

Automatic target recognition generally refers to the use of computer processing to detect and recognize target signatures in sensor data. The conventional ATR architecture comprises a focus of attention (detector and discriminator) followed by a classifier [1]. The role of the focus of attention is to discard image chips that do not contain potential targets. In millimeter wave SAR the focus of attention is traditionally implemented with a simple local intensity test [2] to exploit radar scattering on metallic surfaces. To decrease the number of false alarms a discriminator stage is also built in [1]. The focus of attention works very close to 100% detection rate. All the chips that trigger the focus of attention are further scrutinized by the classifier. ATR classifiers can be broadly divided into two types following the taxonomy in [3]: one class in one network (OCON) and all class in one network (ACON), which is a different nomenclature for the parametric and nonparametric training methodologies described in statistical pattern recognition [25, 30]. As the name indicates, the OCON classifiers are built from each class independently of the others through a statistical description of the class, while the ACON classifiers are developed using the full class set. Hence, ACON classifiers are discriminately trained, which presents some advantages in training efficiency. A landmark example of OCON classifiers in ATR is the template matcher [4]. Template matchers create a class prototype by matching a single class at different pose angles. Therefore one needs many templates per target, and the classifier inplements a linear discriminant function followed by a winner-take-all network (maxnet). It is easy to expand the number of target classes by just developing extra templates. No modification is needed on the developed templates. However, the big problem is that as more templates are added, the more likely the classifier is to make mistakes, in particular when only partial class information is used in the training (as in the matched filter).

The ACON classifier is trained with exemplars of every class using a non-parametric training approach [5]. Classifiers are normally nonlinear, such as the radial basis functions (RBF) or multilayer perceptrons (MLP) which can create universal mappings between the input and the output. The big advantage is that training is discriminant, i.e. one class is trained in the presence of all the other classes. This means that the classifier has to be re-trained from scratch if one more class is added to the problem domain, but the newly trained classifier has the possibility of choosing "features" that best individualize each class with respect to all the others. Although performance also

4

degrades with the number of classes this degradation tends to be slower than with the template matchers.

There is an intermediate class of classifiers where properties of the other classes are brought into the training of an OCON classifier as a penalty term. The minimum average correlation energy (MACE) filter appears as the best example of synthetic discriminant functions (SDF) classifiers [14]. As we commented in a previous paper [6] the MACE is still a compromise between purely template matchers and ACON classifiers, and it is not easy to pick criteria to decide how to best train the MACE. Conventional design "guidelines" are sub-optimal [6].

It is important to analyze in more mathematical terms what is the difference between template matchers and optimal classifiers. The answer is well known in communication theory [7] and statistical pattern recognition [30], and can be simply stated as follows: a template matcher is optimal when all the classes are Gaussian distributed with the same covariance. In fact, assuming Gaussian distributed classes [30] with equal a *priori* probabilities the discriminant function becomes

$$g_i(x) = -0.5(x - u_i)^T \Sigma_i^{-1}(x - u_i) - 0.5 \log|\Sigma_i| + k$$

where u is the class mean, Σ is the covariance matrix. Therefore the general discriminant is a quadratic function. Notice that when the covariance matrix becomes diagonal $\Sigma_i = \sigma^2 I$, the above equation defaults to $g_i(x) = \frac{\|x - u_i\|^2}{2\sigma^2}$, which is a linear discriminant [30]. Linear discriminants are also called distance classifiers because they make decisions based on Euclidean distances from the class mean. Obviously this is not the general case in real world data, so template matchers do not exploit the information contained in the shape of data clusters, which makes them sub-optimal. The optimal classifier for Gaussian distributed classes is the quadratic classifier [55], which is an OCON design. Using quadratic classifiers for SAR-ATR is discouraged due to the large size of the pattern space, and the little data available to train them [25]. Model based ATR as implemented in MSTAR [8, 9] is also intrinsically an OCON design. More sophisticated OCON models can be built based on mixture models [11], or Bayesian principles [10], but the lack of data is the stumbling block. Hence, the issue is how much better can we do for SAR/ATR when the covariance among the targets is exploited indirectly in ACON designs.

We compare here the performance of three different classifier methodologies that have been developed recently in the Computational NeuroEngineering Laboratory (CNEL) at the University

of Florida, or that have not been evaluated in SAR/ATR. One of the CNEL classifiers is an improved template matcher (OCON) that exploits multiresolution of the target signatures. We were interested in evaluating the impact of multi-resolution eigenfeatures in performance. The two ACON classifiers are built on two different philosophies of classification. The SVM classifier proposed by Vapnik nonlinearily projects the data to a generally higher dimensional feature space and sets the linear discriminant function by maximizing the margin; The other ACON classifier projects the data to a smaller dimensional feature space using an information theoretic framework developed at the CNEL to choose the subspace. Here the question is, what is preferable for high performance in SAR/ATR: expansion of features followed by a linear classifier trained for large margin, or projecting the data to a subspace determined to preserve maximally the information for classification? All three classifiers are preceded by a pose estimator to divide the complexity of the task by using the pose of the vehicle. This is also a novel feature in SAR/ATR classifier architecture that has the potential to improve performance and to simplify the computational cost of the testing. We present experimental results obtained on the MSTAR database both for classification and rejection of confusers and end the paper with conclusions.

II. A Classifier Architecture for SAR-ATR

II.1 Pose estimation
The information of the relative position of a target with respect to the sensor, termed the aspect angle of the observation or the pose, is important to decrease the complexity of automatic target recognition ATR. Finding discriminant features among vehicles is simplified if we first choose the target views according to their pose. Effectively we are using a divide-and-conquer strategy to decrease the complexity of the task. However, in SAR-ATR pose estimation is not a simple task due to the enormous variability of the scattering phenomenology among vehicles and across poses. An example of this difficulty is the size of the PEMS (Predict-Extract-Match-Search) module in MSTAR [13], where each target is described by a set of scattering centers. Instead of attempting a model based strategy, we developed a statistical approach for pose estimation, because we believe that statistics is still the most effective way to handle uncertainty and noise in real world phenomena.

In previous papers by the CNEL group [47, 48], a novel pose estimation method was proposed and formulated as the maximization of the mutual information between the aspect angle and the

output of a nonlinear mapper. Generally, pose estimation can be formulated in terms of maximum a posteriori probability (MAP): $\hat{a} = \arg\max_{a} f_{A|X}(a|x) = \arg\max_{a} f_{AX}(a, x)$, where $\hat{a}$ is the estimation of pose a, $f_{A|X}(a|x)$ is the a posteriori probability density function (pdf) given the image x, and $f_{AX}(a, x)$ is the joint pdf of the pose and image. So, the key issue here is to estimate the joint pdf. The very high dimensionality of the image (size: 80x80), however, makes it very difficult to obtain a reliable estimation. Dimensionality reduction (feature extraction) becomes necessary. As shown in Figure 1, the output of the MLP will serve as the feature space for pose estimation $y = MLP(w, x)$ (w are the parameters of the MLP topology - 6400x3x2 or simply a perceptron 6400x2). Hence, instead of working directly on the input image, our pose estimator becomes $\hat{a} = \arg\max_{a} f_{A|Y}(a|y) = \arg\max_{a} f_{AY}(a, y)$. The crucial point for this pose estimation scheme is how well y, which can be interpreted as a feature, conveys information about the pose. To obtain an effective feature, we propose to maximize the mutual information between the feature and the pose $I(y, a)$ as the criterion to train the network so that the feature conveys the most information about pose: $w_{optimal} = \arg\max_{w} I(y= MLP(w, x), a)$. In the results reported here, the joint pdf estimation used for network training produces directly the pose estimation. In [47] a more involved pose estimator (discrete angles) was proposed but the results are identical to the ones obtained with this estimator. The difficulty of this approach is the estimation of the mutual information directly from the data without assumptions on the pdf (information potential field in Figure 1)
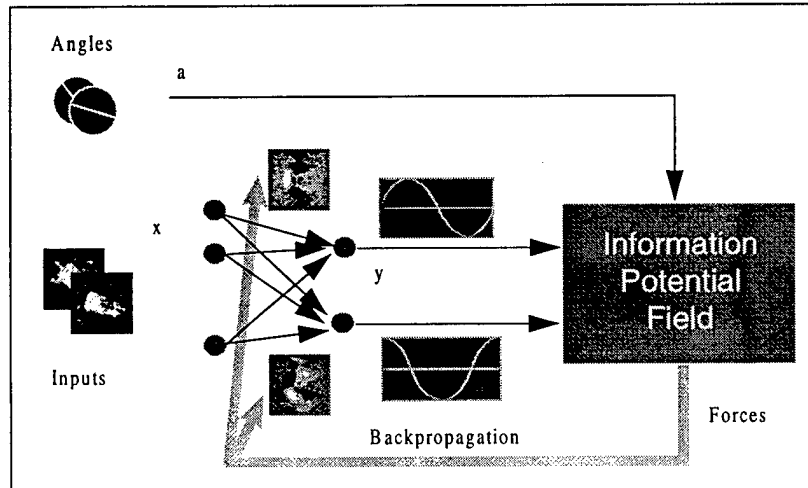

Figure 1. System Diagram

In the appendix we include a summary of our methodology for estimating the mutual information between two data sets, and we will address it later when discussing the ITL classifiers. We con-

7

ducted extensive tests for a one-degree of freedom pose estimation of military vehicles using the MSTAR Database [47,48]. Results are summarized in Table 1. We conclude that when the pose estimator was trained with views of 2 vehicles (T72 and BMP2) at 3.5 degree increments, the pose of other vehicles was estimated with an accuracy better than 5 degrees, and a standard deviation of 3.80. Occlusion tests also showed the robustness of the pose estimator [42]. Another great advantage of this pose estimator is its implementation simplicity. The MLP associates the input with the pose with a simple matrix vector multiplication. Results for a 2-degree of freedom pose estimation task using information theoretic learning are reported in [49].

Table 1: Pose estimation accuracy

vehicle	pose error (standard deviation) in degrees
bmp2_c21_train	1.99 (1.52)
bmp2_c21_test	2.96 (2.41)
t72_132_train	1.97 (1.48)
t72_132_test	3.01 (2.66)
bmp2_9563	2.97 (2.35)
bmp2_9566	3.32 (2.44)
btr70_c71	2.80 (2.33)
t72_s7	3.80 (2.57)

II.2 A novel architecture for SAR-ATR classifiers

The conventional classifier design is based on the matched spatial filter approach, i.e. for each target of interest a template is created through training [4, 12]. The templates are created at 10 degree increments [12] (or less) because correlation, which is the basis of the test, degrades rapidly when there is a pose mismatch between the template and the input image. All the templates are applied to the image chip under analysis and the image chip is classified to the class providing the largest output [12]. These classifiers soon become computational prohibitive for reasonably sized target sets (36 matched filters per degree of freedom). In order to decrease the number of "templates" per class without sacrificing performance, the synthetic [14] or nonlinear discrimi-

nant functions are required because it is known that nonlinear systems normally provide better generalization [36].

The advantage of knowing the pose for classification is that one could divide the task into two stages: first find the pose of the object, and then, select a sub-classifier trained exclusively for that pose range to perform the classification. Figure 2 shows the proposed classifier architecture. First we implement a pose estimator that will select from a bank of classifiers the one that has been trained for that particular pose. Due to the implementation simplicity of the trained pose estimator (a matrix vector product), we further propose to include it in the focus of attention block. In our proposal, the focus of attention will not only choose the image chips for further analysis but will also provide a pose estimation to choose the appropriate classifier.

Here we build 12 classifiers per degree of freedom in the pose, each spanning a 30 degree sector. Since in the MSTAR database the target azimuth is known, and the targets are on the ground plane, this corresponds to a one degree of freedom pose problem and only 12 OCON classifiers per target class are necessary. We can even decrease this number further if we implement ACON classifiers (12 classifiers overall)
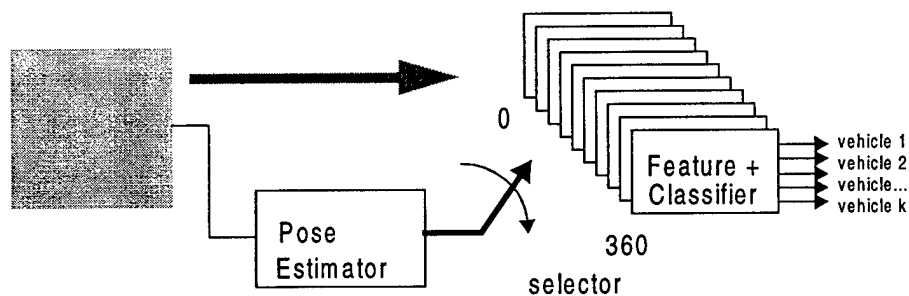

Figure 2 Classifier structure

The results presented in this paper deal with pose sectors of 30 degrees, i.e. for each sector a classifier trained only with the sector data is developed to classify targets. This division is a compromise between the goal of decreasing the overall classifier complexity, the availability of data to train the ACON classifiers, and overall performance. We did not optimize the sector size, but 30 degree sectors were chosen to emphasize the new approach of large angular sectors versus the 10 degree increments utilized in the spatial matched filters. Since our pose estimator has an accuracy of 5 degrees, we can implement this architecture with large confidence (in fact the sectors overlap by 10 degrees in our classifiers to cover the imprecision in the pose).

9

III. Classifiers Design

Two of the main difficulties faced when developing discriminant classifiers for ATR are the lack of training data and how to guarantee generalization. In our opinion, optimal OCON classifiers are out of question for SAR-ATR due to: (1) our lack of knowledge about the statistics of SAR targets. (2) performance penalty by imposing Gaussian assumptions and constraints on the class covariances. Nonparametric classifiers such as K-means [30] require too much data which is lacking in SAR, so they should also be avoided. One promising alternative is the class of ACON classifiers, which are based on artificial neural networks (RBFs or MLPs). These mappers are universal, implementing families of discriminant functions depending on their topologies [15]. MLPs can be readily trained with the backpropagation (BP) algorithm [33], and have been applied to SAR-ATR with reasonable success [16]. However, one difficulty with MLPs trained with the mean square error (MSE) is that this training procedure does not control generalization [20], and more sophisticated training must be implemented. This becomes an issue for SAR-ATR and other similar applications.

In this paper we will investigate improvements on template matchers by creating more features about the target class using multiresolution analysis. We will also compare two different methodologies for ACON classifier design. One of the principles decouples the size of the input space from the number of features by using a Gaussian kernel. The training is based on the concept of maximizing the classification margin as proposed by Vapnik [20]. This SVM classifier has shown very good performance but it has not been extensively tested in SAR/ATR [56]. The other methodology is brand new and projects the data to a subspace such that the projection maximally preserves the information between the desired response and the mapper output [42, 51].

III.1 Multi-resolution decompositions

Detailed discussions of multiresolution are presented in texts on wavelets [17] and pyramidal image processing [18]. While there are many possible reasons for representing an image at several resolutions (or a signal at several scales), we have two main motivations. First, a given feature may be best observed at some scale, but the appropriate scale may not be known. Second, several features may be of interest, but no single resolution is satisfactory for all features. Note that in template based classifiers for SAR-ATR, we utilize only one resolution to describe all the target

features. According to this view, developing templates at several resolutions should improve the performance, because it will be a much more faithful representation of the target signature.

The discussion in this section shows how to decompose a raster scanned image using a cascaded filter approach. The multiresolution components can be organized in a tree structure. One still must choose the basis for decomposition and the appropriate components. We will present PCA-M as an $l2$-energy oriented method for both selecting a basis function and for selecting components.

Figure 3 shows a single decomposition filter (left) and the structure for a tree of $M=3$ levels (middle) and the corresponding synthesis tree (right).
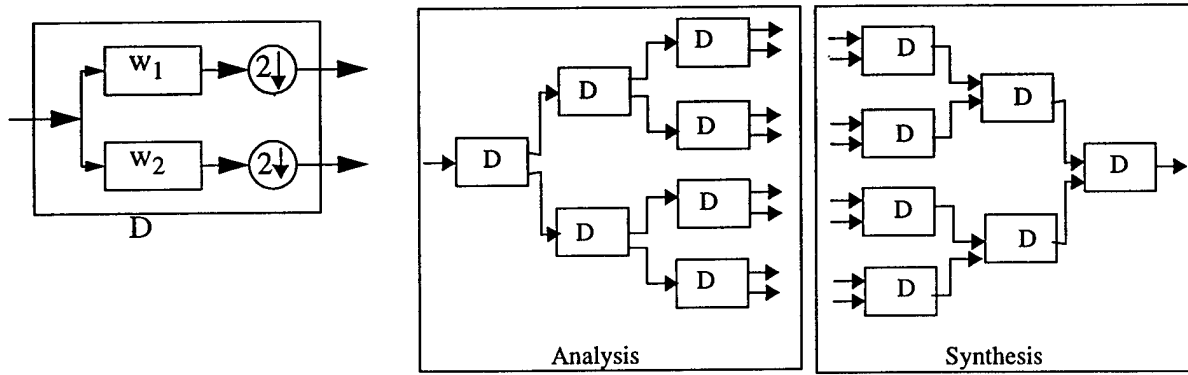

Figure 3. Multiresolution decompositions.

Let us index each signal $x_{m,k}(n)$ such that the first index $m \in [0,...,M]$ denotes the level of decomposition, and the second index $k \in [0,...,m-1]$ identifies one of the signals at that level. With this nomenclature, the original signal is $x(n)=x_{0,0}(n)$, and the signals at the output of the first decomposition filter are $x_{1,0}(n)$ and $x_{1,1}(n)$.

Each level contains a complete "fixed-resolution" representation of the original signal. In this context, completeness implies perfect reconstruction.

$$\begin{bmatrix} x_{m,k}(n) \\ x_{m,k}(n-1) \end{bmatrix} = W \begin{bmatrix} x_{m+1,2k}(n) \\ x_{m+1,2k+1}(n) \end{bmatrix} = WW^T \begin{bmatrix} x_{m,k}(n) \\ x_{m,k}(n-1) \end{bmatrix}$$

When a signal is represented by components at several levels, the representation is said to be multi-resolution or multiscale. A component at a higher level is the output of several cascaded filters. The overall response of the cascade of filters is the convolution of a permutation w_1 or w_2. A

11

higher level signal operates over a longer scale of the input and can take advantage of a richer set of interrelationships. A lower level component has a shorter scale response, but is updated more frequently. The choice of level fixes a trade-off between spatial resolution and scale (richer set of linear combinations of samples of $[x(n),..., x(n - 2m + 1)]$).

Principal Component Analysis (PCA)

PCA and Principal Component Analysis with Multiresolution (PCA-M) are linear transforms,

$$y_k = Tx_k \tag{1}$$

The rows of T form the basis for the output space. Invertibility implies that the mapping preserves all the information needed to (perfectly) reconstruct the original data. When T is invertible, the basis represented by T is said to be biorthogonal. If the inverse of T is its transpose $T^T = T^{-1}$, then the matrix is said to be unitary and the basis is said to be orthogonal.

Given a square matrix A of full rank N, a statement of the eigenvalue problem is $A = W\Lambda W^T$, where Λ is a diagonal matrix of eigenvalues, $\Lambda = diag(\lambda_1, \lambda_2,..., \lambda_N)$, and W is the modal matrix (matrix whose columns are the eigenvectors), $W=[w_1,..., w_k,..., w_N]$. The modal matrix W is unitary and diagonalizes the matrix A.

PCA with Multiresolution (PCA-M)

Although wavelets are complete representations, we would like to choose a basis that will concentrate as much as energy as possible in the lower subspace to provide robust features, that is, features that are resilient to noise. Furthermore, how to choose the most useful tree structure for the decomposition is generally unclear. Due to these facts we propose to integrate PCA decompositions with multi-resolution representations, which we called PCA-M. We propose using energy as the criterion for multiresolution, which we know is optimal for signal representation, and we hope to be highly adequate for classification due to the expansion across scale.

The combined constraints for an orthogonal multiresolution decomposition and principal components analysis can be simultaneously met by using the iterated filter approach described in the above section, but they are not adaptive (Haar basis) [19].

One of our desired capabilities for PCA-M was to extract highly compressed high-energy components first, because it is more robust with respect to scaling shifts and produce high SNR features. This goal forced us to sacrifice orthogonality among different resolution components, because, we

felt that orthogonality was less critical for classification than for reconstruction. The non-orthogonal decomposition can be implemented with a partially connected network (Figure 4) using the Generalized Hebbian Algorithm [15]. The GHA operates directly on the training set images, being trained one image at a time.
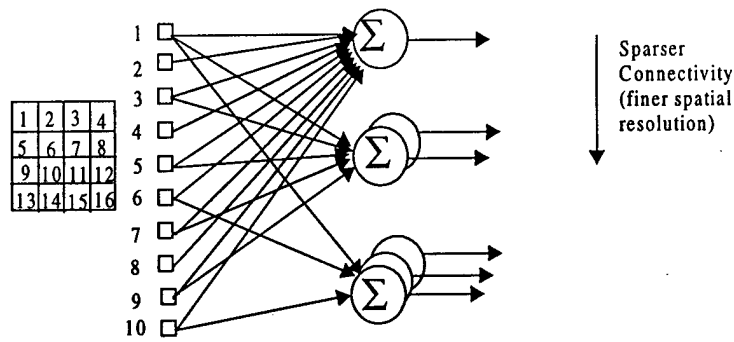

Figure 4. Multi-resolution PCA

Eigenvectors at a given resolution will be ordered by energy and are orthogonal. Eigenvectors across different resolutions will not be orthogonal or necessarily ordered by energy. Again, the reason for our approach is to accept loss of spatial resolution as long as there are high-energy components to extract. After the highest energy components are identified, we seek to increase spatial resolution by moving to a lower dimensional subspace.

A BMP2 Image and Components

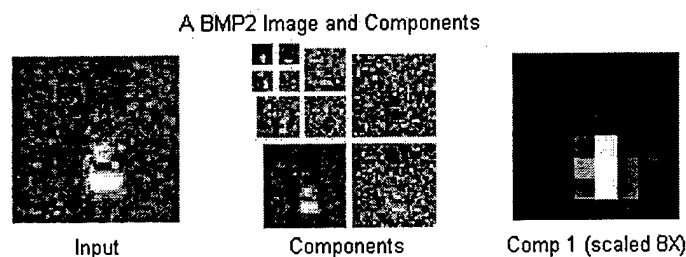Input   Components   Comp 1 (scaled 8X)

Figure 5 A BMP2 image and components

A sample decomposition (middle) and a close-up of the first component (right) are shown in Figure 5. The middle display shows four level 3 components (approximation + 3 detail images), three level 2 components (16×16 detail images), and three level 1 (32×32 detail images). The number of components at each level is motivated only for comparison with approximation and detail signals in wavelet multiresolution analysis. The number of levels was chosen based on performance in other applications [19].

The classifier architecture is a OCON network followed by a maxnet (to find the maximum).

- **Classifier Structure** - The overall classifier is a parallel structure of 10 individual template matchers per class. Each individual classifier operates on a single component of the input image, and since there are three target classes, there are three outputs. The output with the highest value corresponds to an intermediate classification of the image based on the given component. The final classification is based on a simple majority vote among individual classifiers.

- **Classifier Weights** - The network weights (templates) are the normalized averages for the corresponding components of the training set images, e.g., the weights of the connections to the output corresponding to the BMP-2 class are obtained by averaging the first components of all the BMP-2 images. Hence we are implementing matched filters at different resolutions.

- **Overall Classification** - The final classification is based on a majority of equally weighted votes among individual classifiers. A minimum number of votes can be used to set a threshold for rejecting an image (detection) but here the outputs of the parallel networks are summed so that classification and rejection are done only at this final stage.

III.2 Support Vector Machines

The perceptron or MLP trained with the error back-propagation implements the empirical risk minimization, because it only takes into consideration the performance in the training set. However, as indicated in [22], neither the perceptron criterion nor the MSE criterion would necessarily lead to a minimum classification error in the test set, i.e. they do not guarantee good generalization. In this section, a learning criterion for structural risk minimization [20] is considered. The advantage of a SVM classifier is that it can decouple the number of free parameters of the learning machine from the input space dimensionality [24].

The Optimal Hyperplane

The training set is said to be separated by an optimal hyperplane (OH) if the following two conditions are satisfied. First, all the samples are separated without error (keep the empirical risk zero), and second, as illustrated in Figure 6, the distances between the closest vectors to the hyperplane are maximal. The separating hyperplane is described in the canonical form, i.e.,

$$y_i(w \cdot x_i + b) \geq 1, \ i \ = \ 1, \ ..., m \tag{2}$$

It is easy to prove that the margin between the two hyperplanes $H_1 : \mathbf{w} \cdot \mathbf{x}_i + b = 1$ and $H_2 : \mathbf{w} \cdot \mathbf{x}_i + b = -1$ is $d = 2/\|\mathbf{w}\|$. Thus, to find a hyperplane that satisfies the second condition, one has to solve the quadratic programming problem of minimizing $\|\mathbf{w}\|^2$, subject to constraint (2).
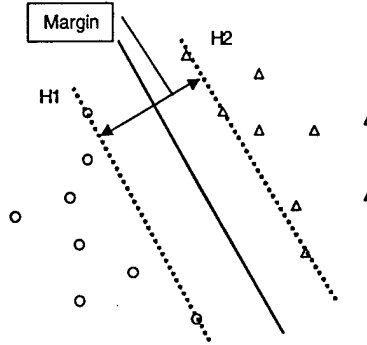

Figure 6. A two-class linearly separable problem (balls vs. triangles). The optimal hyperplane (solid line) intersects itself halfway between the two classes, and keeps the margin maximal. The samples across the boundary *H1* or *H2* are support vectors.

The solution to this optimization problem is given by the saddle point of a primal Lagrange functional,

$$L_p \ = \ \frac{1}{2}\|w\|^2 - \sum_{i=1}^{m} \alpha_i [y_i(w \cdot x_i + b) - 1] \tag{3}$$

where $\alpha_i, \ i = 1,...,m$, are positive Lagrange multipliers. Since (3) is a convex quadratic programming problem, this means that it is equivalent to solve a "dual" problem [23]: maximize L_p, subject to the constraints that the gradient of L_p with respect to *w* and b vanish, which gives the conditions

$$w = \sum_i \alpha_i y_i x_i$$

$$\sum_i \alpha_i y_i = 0 \tag{4}$$

Substituting (4) into (3), we get the dual problem of maximizing,

$$L_D = \Lambda^T \cdot 1 - \frac{1}{2}\Lambda^T C \Lambda$$

$$\Lambda^T Y = 0 \tag{5}$$

$$\Lambda \geq 0$$

where $\Lambda^T = (\alpha_1, \alpha_2, \ldots \alpha_m)$ is a parameter vector, $1^T = (1, \ldots m)$ is an m-dimensional unit vector, $Y^T = (y_1, \ldots, y_m)$ is the m-dimensional label vector, and C is a symmetric m by m correlation matrix with elements $C_{ij} = y_i y_j x_i \cdot x_j, i, j = 1, \ldots, m$. Notice that there is a Lagrange multiplier α_i for every training sample. In the solution, those points for which $\alpha_i > 0$ are called "support vectors" (SV), and lie on either H_1 or H_2. The separating rule is, based on the Optimal Hyperplane,

$$g(x) = \mathrm{sgn}\left(\sum_{i \in SV} y_i \alpha_i x \cdot x_i + b \right) \tag{6}$$

Kernel Based Classifiers and SVMs

Until now, all the previous architectures create the decision functions that are all linear functions of data. Then one may ask how can the above method be generalized to the case of a nonlinear decision functions? One alternative is to map the data to some high dimensional feature space using a mapping $\phi : R^d \to E$. There is evidence provided by Cover's theorem [29] that a complex pattern classification problem nonlinearly mapped onto a high-dimensional space is more likely to be linearly separable than the original low-dimensional space. The advantage of this method is that it decouples the numbers of free parameters of the learning machine from the input space dimensionality. In this way, the decision rule of (6) is implemented in the new feature space, i.e.,

$$g(x) = \text{sgn}\left(\sum_{i \in SV} y_i \alpha_i \phi(x) \cdot \phi(x_i) + b\right)$$

(7)

By the Mercer's condition, there exists a mapping ϕ and a symmetric function $K(x,y)$ which has an expansion $K(x, y) = \sum_{k=1}^{\infty} \phi_k(x)\phi_k(y)$, if and only if, for any $f(x)$ such that $\int f^2(x)dx$ is finite, there exists,

$$\int K(x, y)f(x)f(y)dxdy \geq 0$$

(8)

The convolution of the inner product allows the construction of a decision function that is nonlinear in the input space,

$$g(x) = \text{sgn}\left(\sum_{i \in SV} y_i \alpha_i K(x_i, x) + b\right)$$

(9)

and this is also equivalent to a linear decision function in the high-dimensional feature space of $\phi_1(x), \ldots, \phi_m(x)$. This learning machine is the so-called Support Vector Machine.

Training SVMs with the Adatron Algorithm

The so-called Adatron algorithm [26] was proposed to solve the quadratic programming problem using the concept of adaptive learning. The basic idea of the Adatron algorithm is that, instead of updating the weight w directly, one can update the Lagrange multipliers. Since the weight vector now resides in the feature space defined by the kernels, they can not be accessed for direct update. However, the multipliers are still accessible in the input space and can be updated. The Adatron performs a gradient descent in the quadratic risk function. At each step the gradient of the risk function is computed in the direction of one canonical basis vector, then the solution is updated. In general, the Adatron algorithm implements a form of gradient descent in the convex risk function. This fact was investigated in detail in the kernel Adatron with bias and soft-margin algorithm

[27]. Compared with the numerical quadratic programming method, the kernel Adatron algorithm can learn large margin decision functions in kernel feature space in an iterative "on-line" fashion. The requirement is to store in matrix form the full input data. In the case of high dimensional input space and few data samples, this storage is reasonable and it saves computations when compared to the quadratic programming solution.

III.3 Information-Theoretic Criteria: Supervised Learning with Quadratic Mutual Information

Learning and adaptation are intrinsically related to information theory [32, 35]. In general, a learning rule should make full use of the information which is available in the data while avoiding the use of any extraneous information directly or indirectly imposed by the solution. This is the often discredited Laplace principle of insufficient reasoning, which was reformulated in Jeffreys' uninformative priors [34], and finally scientifically formulated by Jaynes as the maximum entropy (MaxEnt) principle [43].

Supervised learning utilizes two sources of information to train the learning machine: the input data and the class labels, which are known in a training set. In classification, the output of a mapping should convey as much information as possible about the input with respect to the class labels. The mutual information principle provides the answer. In the following, the basic idea of entropy and mutual information will be reviewed briefly and how to train an MLP with MI will be addressed.

Quadratic Entropy and Mutual Information

Shannon definition of information [37] is $H_S(Y) = \int p(y)\log\frac{1}{p(y)}dy$, where $p(y)$ is the probability density function of the random variable Y. There are many more definitions of entropy using the theory of means. In particular, Renyi's entropy with order α, which we will denote by $H_{R\alpha}$ is defined as [38, 39]

$$H_{R\alpha} = \frac{1}{1-\alpha}\log\left(\int_{-\infty}^{+\infty} f_Y(y)^{\alpha}dy\right) \qquad \alpha > 0, \alpha \neq 1 \qquad (10)$$

When $\alpha = 2$, $H_{R2}(Y) = -\log\left(\int_{-\infty}^{+\infty} f_Y(y)^2 dy\right)$ is called Quadratic entropy for convenience. In

the appendix we describe a procedure to estimate quadratic entropy directly from data samples as

$$H(\{a_i\}) = H_{R2}(Y|\{a_i\}) = -\log\left(\int_{-\infty}^{+\infty} f_Y(y)^2 dy\right) = -\log V(\{a_i\})$$

$$V(\{a_i\}) = \frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}\int_{-\infty}^{+\infty} G(y-a_i, \sigma^2)G(y-a_j, \sigma^2)dy = \frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N} G(a_i - a_j, 2\sigma^2)$$

where G(.) is a multidimensional symmetric Gaussian kernel. Effectively we are estimating quadratic entropy by using a nonparametric estimator based on the integral of the pdf estimated with Parzen windows [40]. *Note that this is a very interesting expression since it tells that quadratic entropy is associated with interactions among pairs of data samples a_i and a_j.* The general form of $V(\{a_i\})$ is a potential field, which we called an *information potential.* Maximizing entropy is equivalent to minimizing the information potential. But how can we manipulate information on a set of samples? We have to interpret the samples as outputs of mapper $y_i=a_i$ [42]. Since the outputs are a function of the system parameters $y = f(x,w)$, changing the parameters of the system will change the relative position of the outputs in the output space, modifying the entropy of the set. With another analogy from physics, the derivative of the potential field is a force, so if we take the derivative of $V(Y)$ w.r.t. a_i we have an *information force* on sample i given by [42],

$\varepsilon_i = \frac{\partial}{\partial a_i}V(\{a_i\})$. This can be interpreted as an error signal that can be *incorporated in the back-propagation algorithm* [15] to change the MLP weights such that the entropy of the set is maximized (or minimized) in the output space. *This provides a new, unsupervised principle to train MLPs with information theoretic criterion, which we called Information Theoretic Learning (ITL).*

Still another information measure useful to quantify the entropy between pairs of random variables is mutual information. The mutual information between two variables Y_1 and Y_2 is the Kullback-Leibler divergence [41] between the joint pdf and the factorized marginal pdf:

$$I(Y_1, Y_2) = K(f_{Y_1Y_2}(y_1, y_2), f_{Y_1}(y_1)f_{Y_2}(y_2)) = \int\int f_{Y_1Y_2}(y_1, y_2)\log\frac{f_{Y_1Y_2}(y_1, y_2)}{f_{Y_1}(y_1)f_{Y_2}(y_2)}dy_1 dy_2 \quad (11)$$

where $f_{Y_1 Y_2}(y_1, y_2)$ is the joint pdf, $f_{Y_1}(y_1)$ and $f_{Y_2}(y_2)$ are marginal pdfs. Kullback-Leibler divergence between two pdf $f(x)$ and $g(x)$ is defined as:

$$K(f, g) = \int f(x) \log \frac{f(x)}{g(x)} dx \tag{12}$$

where implicitly the Shannon's entropy is utilized. From (10) we can observe that unfortunately it is not quadratic in the pdf so it can not be easily integrated with the Parzen window pdf estimation. Therefore, a new divergence or distance measure between two pdfs which contains only quadratic terms is needed. Based on the Euclidean difference of vectors inequality we can write

$$\|x\|^2 + \|y\|^2 - 2x^T y \geq 0 \tag{13}$$

Hence, we can propose to estimate the divergence between two PDFs $f(x)$ and $g(x)$ based on the Euclidean distance as

$$I_{ED}(f, g) = \int f(x)^2 dx + \int g(x)^2 dx - 2 \int f(x) g(x) dx \tag{14}$$

It is easy to show that $I_{ED}(f, g) \geq 0$ and the equality holds true if and only if $f(x) = g(x)$ ($\int f(x) dx = \int g(x) dx = 1$), so it is a definition for a distance between two pdfs, which we call Euclidean Distance Quadratic Mutual Information (ED-QMI). We have examined the similarity of $I_{ED}(f, g)$ with mutual information in a number of cases, and we conclude that they display the same minima and maxima [51]. An added benefit is that $I_{ED}(f, g)$ can be computed from the information potential, so we have a procedure to estimate mutual information directly from samples.

Supervised training of the MLP with quadratic mutual information
We can utilize the idea of Euclidean distance (ED) to express ED-QMI as

$$I_{ED}(Y_1, Y_2) = \left(\iint f_{Y_1 Y_2}(z_1, z_2)^2 dz_1 dz_2 \right) + \left(\iint f_{Y_1}(z_1)^2 f_{Y_2}(z_2)^2 dz_1 dz_2 \right) - 2 \left(\iint f_{Y_1 Y_2}(z_1, z_2) f_{Y_1}(z_1) f_{Y_2}(z_2) dz_1 dz_2 \right) \tag{15}$$

Basically (15) measures the Euclidean distance between the joint pdf and the factorized marginals. With the definitions in the appendix it is not difficult to obtain

$$I_{ED}((Y_1, Y_2)|y) = V_{ED}(y)$$

$$V_{ED}(y) = \frac{1}{N^2} \sum_{i=1}^{N} \sum_{j=1}^{N} V_{ij}^1 V_{ij}^2 - \frac{2}{N} \sum_{i=1}^{N} V_i^1 V_i^2 + V^1 V^2 \tag{16}$$

where V_{ij} stands for the information potential in the joint field, and V_i for the marginal potential. It is also not difficult to obtain the formula for the calculation of the information force produced by the cross information potential field V_{ED} as

$$c_{ij}^k = V_{ij}^k - V_i^k - V_j^k + V^k, \quad k = 1, 2$$

$$F_i^l = \frac{\partial V_{ED}}{\partial y_i^l} = \frac{-1}{N^2 \sigma^2} \sum_{j=1}^N c_{ij}^k V_{ij}^l d_{ij}^l \tag{17}$$

$$i = 1, ..., N, \quad l \neq k, \quad l = 1, 2$$

where c_{ij}^k are cross matrices which serve as force modifiers.

In a supervised framework we would like to maximize the information between the class label and the output of the MLP. So the joint space (Y_1, Y_2) is the space of the classes (c_i) versus the MLP outputs (y_i), i.e. $Y_1 = \{c_1, ..., c_L\}$ and $Y_2 = \{y_1, ... y_p\}$. Note that the class assignments are solely used to estimate the marginal information potentials and the cross-information potential, so they just provide a way of dividing the data. They do not provide numerical targets, which is in tune with the a priori information available.

Once the MLP is trained, we can implement a classifier in different ways: either we use the training set MLP outputs to design a Bayes classifier by estimating the mean and covariances per class, or we use again a Parzen estimator to estimate the likelihoods per class, or finally we can train another MLP that uses the maximum information projections as the pattern space. Since the pattern space is normally of much smaller dimension (typically 2 to 6 in our studies) any of the methods works well. We will be testing the likelihood method using the Parzen estimator.

IV. Results and Discussion

Here we will present classification and recognition results obtained with the three classifiers presented in Section III: a linear classifier using principal component analysis with multiresolution as the frontend (PCA-M), a support vector machine (SVM), and a classifier trained with quadratic mutual information (QMI).

In this paper, synthetic aperture radar automatic target recognition experiments were performed using the MSTAR database to classify three targets. The data are 80 by 80 SAR images drawn from three types of ground vehicles: the T72, BTR70, and BMP2 as shown in Figure 7. These

images are a subset of the 9/95 MSTAR (Moving and Stationary Target Acquisition and Recognition) Public Release Data, where the pose (aspect angles) of the vehicles lies between 0 to 360 degrees. Only target images are used here (there is no need for the focus of attention) so they will be directly scored by the classifier. The classifier includes 12 sector classifiers (30 degree sectors). Our results assume that the pose estimator is error free because during testing the data is presented to the appropriate sector classifier.

We normalize to one the L2-norm of all the images from the training and testing sets. This preprocessing was kept at a minimum because the targets in the MSTAR database were in the same open field background, and the radar was carefully calibrated. The target chips were used directly without re-centering nor masking of background to individualize the targets. If these operations were performed better accuracy should be possible, but a longer effort would have been necessary to conduct the testing.

The training set contained SAR images taken at a depression angle of 17 degrees, while the testing set depression angle is 15 degrees. Therefore the SAR images between the training and the testing sets for the same vehicle at the same pose are different, which helps to test the classifier generalization. Variants (different serial number) of the three targets were also used in the testing set, as illustrated in Table 2. The size of training and testing sets is 698 and 1365, respectively.

Table 2: Training and Testing Set

Training Set	Size	Testing Set	Size
T72(Sn_132)	232	T72(Sn_132)	196
		T72(Sn_812)	195
		T72(Sn_s7)	191
BTR70(Sn_c71)	233	BTR70(Sn_c71)	196
BMP2(Sn_c21)	233	BMP2(Sn_c9563)	195
		BMP2(Sn_c9566)	196
		BMP2(Sn_c21)	196

Each classifier was trained at its full potential. The QMI employs a single-layer perceptron with 80x80 input nodes and 3 output nodes. The SVM uses a RBF network and the three class classifier is obtained by training in a pairwise fashion. The PCA multiresolution uses a 64x64 input field analyzed by a set of 10 matched filters for each of the higher energy components of the three

scales. The learning rates, and other free parameters in each method were fine tuned for best performance. Although the classifiers were independently developed, they are using the same training and test data and the same problem specification, so their outputs can be directly compared.
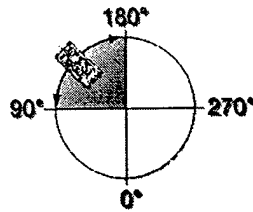
(a)
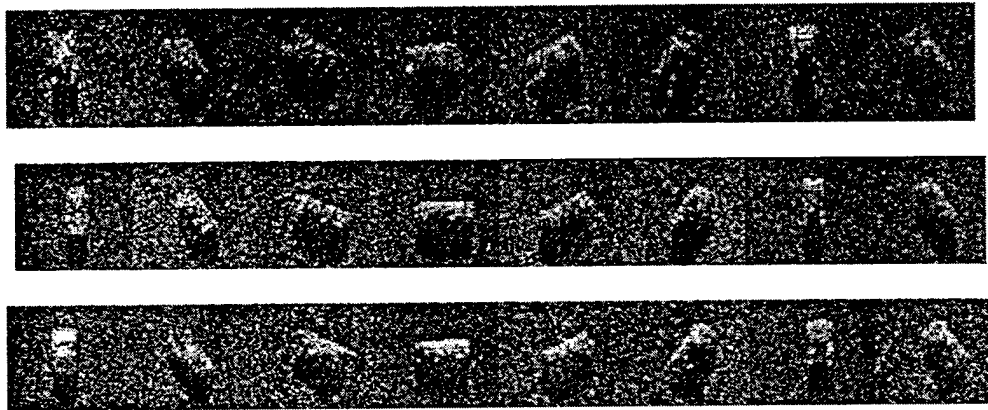
(b)

Figure 7. (a) Target pose (b) Example of the data set (BMP2, T72, BTR70)

Classification Results

The base line for the comparison is the template matching method [12], presented in Table 3. Reference [12] describes a power normalized template matcher developed with templates at 10 degree increments and a mask individualizing the targets, but using the same MSTAR target mix. The classification results of our three classifiers are summarized by confusion matrices through Table 4 to 6. For all the four classifiers, a threshold was set for each method to keep the probability of detection[1], Pd, equal to 0.9 in the testing set (A Pd of 0.9 is typically used in MSTAR and

---

1. Probability of Detection (Pd) is defined here as number of targets detected / number of targets tested.

recommended as a standard operating point). These tables present the actual counts per class, and an overall classification performance is give in Table 7.

Table 3: Template Matching

	BMP2	T72	BTR70
BMP2	483	9	59
T72	43	427	16
BTR70	4	0	188

Table 4: PCA-M

	BMP2	T72	BTR70
BMP2	487	29	19
T72	66	441	25
BTR70	1	0	183

Table 5: QMI

	BMP2	T72	BTR70
BMP2	509	12	4
T72	38	441	26
BTR70	1	0	192

Table 6: SVM

	BMP2	T72	BTR70
BMP2	511	15	14
T72	31	453	10
BTR70	0	0	195

From the overall classification results in Table 7, we conclude that the multiresolution template matcher (PCA-M) performance is slightly worse than the standard template matcher [12]. But the difference here is that we do not use a template matcher every 10 degrees of pose (aspect angle), but one every 30 degrees. This shows that using templates at different scales preserves classifica-

24

tion accuracy and is able to tolerate larger errors induced by pose estimation. For the two ACON classifiers, the SVM gets the best classification accuracy, with a misclassification error of 5.13%, closely followed by the QMI with a miscalssification of 5.93%. The two ACON classifiers are better than either of the two template matchers. The reason has been attributed in Section II and III to both nonlinear discriminant functions and discriminant training.

Table 7: Overall misclassification error

Classifier	Error Rate
Template	9.60%
PCA-M	10.3%
QMI	5.93%
SVM	5.13%

While we were performing the ATR experiments, another group presented results with a SVM. In [56], a SVM classifier was used to classify the same target mix in MSTAR, but using a polynomial instead of a Gaussian kernel function. The reported misclassification errors are around 6.6%-7.2%, slightly worse than our results.

Recognition and Confuser Rejection

A critical problem in ATR is how to discriminate between target and non-target vehicles, the so-called confusers. When we cannot guarantee that all the vehicles found in the test set belong to the training set classes, rejecting patterns with a low degree of membership to these classes becomes important. In this experiment two non-target vehicles (confusers), D7 and 2S1, were added to the testing set. In a sense, the confusers are especially hard cases of false alarms actually leaking into the classifier. The size of both confuser sets is 274.

The rejection results are listed in Table 8. From the table, we conclude that our three classifiers give better results of confuser rejection than the template matcher. Among them, the SVM gets the best result, rejecting more than two third of all the confusers, while the standard template method rejects 53.5% of confusers. This good performance of SVMs is attributed to the fact that the Gaussian kernel function implements a local discriminant in feature space that tends to represent better the class. The MLP uses an intersection of hyperplanes which are global discriminants

and tend to provide decision surfaces that do not represent well the clusters. This may explain the reason why the QMI rejection is below that of the PCA-M.

Table 8: Confuser Rejection

| Classifier	Rejection
Template	53.5%
PCA-M	60.0%
QMI	54.5%
SVM	68.8%

To give an overall performance comparison among the classifiers tested, the receiver operating characteristics (ROC) curves of the two ACON classifiers are shown in Figure 8-10. There are two kinds of ROC curves of interest: one is Probability of detection (Pd) vs Probability of false alarm (Pfa[1]), and the other is Probability of correct classification (Pcc[2]) vs Pfa.

---

1. Probability of false alarm (Pfa) is defined as the probability that a specific set of confusers will be detected as targets, i.e., number of confusers detected / number of confusers tested.
2. Probability of correct class (Pcc) is defined as number of targets correctly classified / number of targets tested.
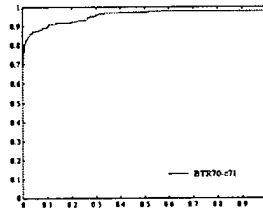
Composite Detection ROC Curve
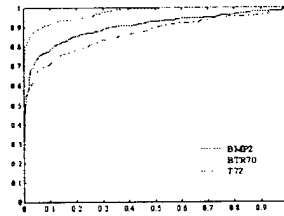
BMP2 Classification ROC Curve
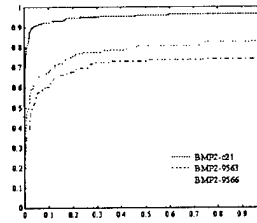
BTR70 Classification ROC Curve
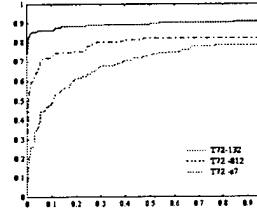
T72 Classification ROC Curve

Figure 8 Template matching: Pd vs Pfa ROC curve (upper left) and Pcc vs Pfa ROC
curve. (from website: //www.standevalexp.vdl-atr.afrl.af.mil/New_MSE_Results)

The ROC curves for the PCA-M are not easy to construct because this classifier was built from
several independent classifiers for each resolution. Hence, multi-dimensional ROCs would be
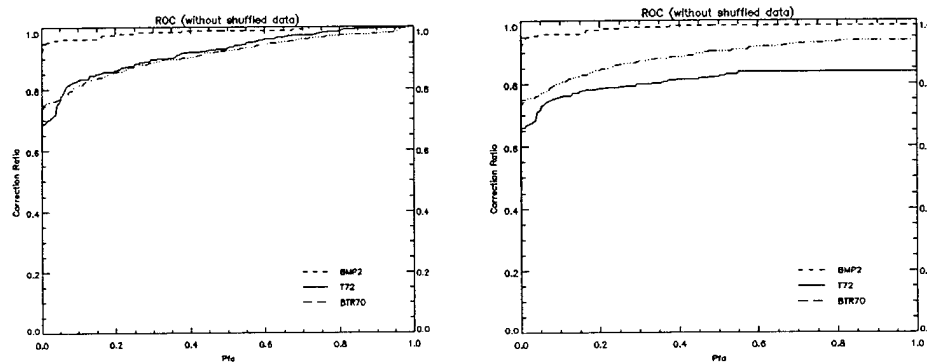needed to fully characterize the performance of the PCA-M classifier.

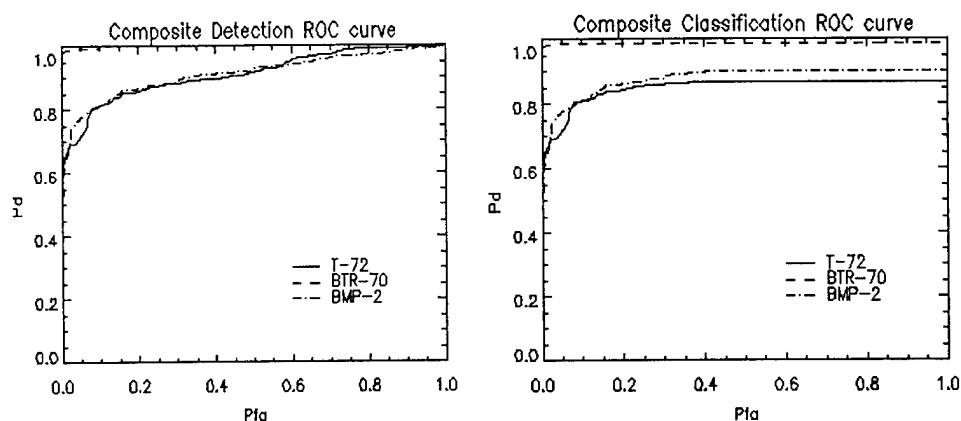Figure 9 QMI: Pd vs Pfa ROC curve (left) and Pcc vs Pfa ROC curve (right)

Figure 10 SVM: Pd vs Pfa ROC curve (left) and Pcc vs Pfa ROC curve (right)

One aspect to point out is that the ROC curve for the BTR70 is much better than the other two, because the training and testing sets of BTR70 are from the same serial number (but at different depression angle), while the testing sets of the other two targets are different variants from the training sets (see Table 2).

Another interesting observation is that the SVM and QMI present very similar ROC curves. Actually, the two methods employ learning mechanism that are not as different as they may seem at first. Although the SVM tries to minimize the confidence interval and maximize the classification margin and the QMI maximizes the mutual information between class labels and classifier outputs, they share one common point. They try to extract as much information as possible from pairs of training set samples by either using the kernel correlation matrix (SVM) or the cross-information potential (QMI), which represent high order statistics information instead of the second order statistics (Euclidean distance) used in the template matcher.

The shape of the discriminants is dictated by the learning machine topology and plays a major role in rejection of confusers. The multilayer perceptron utilizes hyperplanes (global discriminants) while the SVM with Gaussian kernels uses local discriminants. In recognition applications, a critical problem is how to protect the classifier against any potential confuser. One possibility to define rejection is to quantify the degree of "membership" to the class, but it is very difficult to do so since the true underlying probability density function is not available. A simple proxy is obtained by thresholding the output of the classifier, which substitutes the pdf information by the

28

class discriminant. This is what we implicitly do when varying the threshold to plot the ROC curve. However, for meaningful results, class discriminants should be confined to the local area of the pattern space where the samples are located, that is, the discriminants should be local. The confuser rejection results of Table 8 showed that the SVM with Gaussian kernel, which implements a bounded "local" decision region in the input space, in fact obtains the best confuser rejection. The SVM maps a confuser far away from the "local" decision region onto a location close to the origin of the feature space, which promises a reliable rejection. The template matcher is based on a distance to the center of the cluster so it is also local. However, the QMI is here used to train a MLP that possesses global discriminants (QMI can also train RBFs, but this topology was not used here). Hence, although the MLP trained with QMI provides better classification performance than the PCA-M classifier, it ranked under the multiresolution template matcher in confuser rejection. This result suggests that the decisive factor for high rejection to confusers lies more in the neural network topology than on the training algorithm.

V. Conclusions

Our work proposes a novel architecture for SAR ATR that includes the pose estimator in the focus of attention block. Our pose estimator based on information theoretic principles is accurate within 5 degrees in MSTAR data, and can be efficiently implemented with a vector matrix multiply. This is one of the advantages of adaptive systems in frontend ATR. The time consuming step is the training of the adaptive system, which is done off-line. Once the system is trained, the information about the domain is stored in the adaptive weights, and a test can be done very fast. Hence, the natural place to include the pose estimator is in the focus of attention block. This block will not only flag image chips that are potential targets, but also will provide the pose. This concept can be further explored to also classify man-made clutter and further improve the rejection of ATR systems.

We investigated here classifier designs for large pose sectors with the goal of decreasing the computational complexity of template matchers without affecting (and eventually improving) classification performance and rejection to confusers. ACON classifiers yield one classifier per sector independently of the number of classes, so they are the most efficient design. When implemented with nonlinear topologies they should be able to implement accurate classifiers. Still we would

like to see if template matchers could handle large pose sectors, so we developed a new multi-resolution template matcher based on multi-resolution PCA.

We demonstrated the efficiency of our proposed classifier structure and the three methods of learning and representation with a target mix from the MSTAR data set. An important result is the excellent performance of our proposed QMI and SVM classifiers. The QMI chooses the projection from the input space to the output by maximizing the mutual information between the desired response (the labels) and the output of the classifier, while the SVMs de-couples the numbers of free parameters of the learning machine from the input space dimensionality. Either method provides high performance classifiers. In detection the higher performance of the SVM classifier is attributed not to the training but to the local nature of the discriminant functions obtained with the Gaussian kernels. In a different paper [53] we showed that the optimal hyperplane (a perceptron trained with the Adatron algorithm), performs in detection at the same level as the perceptron trained with ITL. Hence, we conclude that for detection, classifier topologies should be built with local discriminant functions such as the Gaussian.

The PCA-M classifier is an intuitively appealing classifier structure, because it works with features at several spatial scales. Here we only experimented with 3 scales, and utilized a very simple equal weight voting strategy as input to the maxnet to choose the class. This scheme is too naive and can be easily improved with a linear or nonlinear network that learns how to weight the different contributions from each scale for best performance. Another issue that needs a more systematic treatment is the selection of thresholds for each scale. Nevertheless, this classifier shows that multi-scale templates do provide good rejection and reasonable classification accuracy for sectors as large as 30 degrees, which would be disastrous for the conventional template matcher. More work needs to be done to improve the PCA-M.

References

[1] J. Principe, A. Radisavljevic, J. Fisher, and L. Novak. "Target prescreening based on a quadratic Gamma discriminator". *IEEE Transactions on Aerospace and Electronic Systems*, 34(3), 706-715 (1998).

[2] L. Novak, G. Owirka, W. Brower and A. Weaver. "The automatic target recognition system in SAIP", *The Lincoln Lab Journal*, 10(2), 187-202 (1997).

[3] S. H. Lin, S. Y. Kung, and L.J. Lin. "Face recognition/detection by probabilistic decision based neural network". *IEEE Transactions on Neural Networks*, 8(1), 114-132 (1997).

[4] A. Vander Lugt, "Signal detection by complex matched spacial filtering", *IEEE Trans. Information Theory*, 10(23), (1964).

[5] C. Bishop, *Neural Networks for Pattern Recognition*, Clarendon Press, Oxford, (1995).

[6] J. Fisher and J. Principe "Recent advances to nonlinear MACE filters", *Optical Engineering.* 36(10), 2697-2709 (1998).

[7] H. L. Van Trees, *Satellite Communications*, Wiley, 1979.

[8] E. Keydel, et al., "Signature prediction for model-based automatic target recognition", *Algorithms for Synthetic Aperture Radar III*, E. Zelnio and R. Douglass, Eds., Proceedings of the SPIE, Vol. 2757 (1996).

[9] E. Keydel, et al., "Reasoning support and uncertainty prediction in model-based vision ATR", *Algorithms for Synthetic Aperture Radar Imagery VI*, E. Zelnio, Eds., Proceedings of the SPIE, Vol. 3721. (1999)

[10] J. Ruanaidh and W. Fitzgerald, *Numerical Bayesian methods applied to signal processing*, Springer Verlag, (1996)

[11] R. A. Jacobs, M. I. Jordan, S. J. Nowlan, and G. E. Hinton, "Adaptive mixtures of local experts", *Neural Computation*, 3, 79-87 (1991).

[12] T. Ross, V. Velten, J. Mossing, S. Worrell, and M. Bryant. "Standard SAR ATR Evaluation Experiments using the MSTAR Public Release Data Set". *Algorithms for Synthetic Aperture Radar Imagery V*, E. Zelnio, Eds., Proceedings of the SPIE, Vol. 3370, 566-573. (1998)

[13] J. Diemunsch and J. Wissinger. "MSTAR model-based automatic target recognition: search technology for a robust ATR". *Algorithms for Synthetic Aperture Radar Imagery V*, E. Zelnio, Eds., Proceedings of SPIE, Vol. 3370, 481-492. (1998)

[14] C. F. Hester and D. Casasent, "Multivariant technique for multiclass pattern recognition", *Appl. Opt.* 19, 1758-1761, (1980).

[15] Simon Haykin, *Neural Networks, A Comprehensive Foundation*, Macmillan Publishing

Company (1994).

[16] S. Rogers, et al., "Neural networks for automatic target recognition", *Neural Networks*, 8, 1153-1184, (1995)

[17] M. Vetterli and J. Kovacevic. *Wavelets and Sub-band Coding*. Prentice-Hall, Inc., Englewood Cliffs, NJ (1995).

[18] Anil K. Jain. *Fundamentals of Digital Image Processing*. Prentice-Hall, Englewood Cliffs, NJ, (1989).

[19] V. Brennan and J. Principe, "Face classification using PCA and multiresolution", *Proc. IEEE Workshop on NNSP*, 506-515, (1998).

[20] V. Vapnik. *The nature of statistical learning theory*. New York: Springer-Verlag, Inc., (1995).

[21] F. Rosenblatt. "The Perceptron: A probabilistic model for information storage and organization in the brain", *Psychological Review* 65, 386-408, (1958).

[22] B. Juang and S. Katagiri. "Discriminative learning for minimum error classification." *IEEE Transactions on Signal Processing*, 40(12): 3043-3054, (1992).

[23] R. Courant and D. Hilbert. *Methods of mathematical physics*, Interscience, (1953).

[24] C. Burges. "A tutorial on support vector machines for pattern recognition". *Data Mining and Knowledge Discovery*, (1998).

[25] K. Fukunaga. *Statistical pattern recognition*. 2nd ed. San Diego, CA:Academic Press.

[26] J. Anlauf and M. Biehl. "The Adatron: an adaptive perceptron algorithm". *Europhysics Letters*, 10(7), 687-692, (1989).

[27] T. Frieb, N. Cristianini and C. Campbell. "The kernel-Adatron algorithm: a fast and simple learning procedure for support vector machines". *Machine Learning: Proceedinds of the 15th International Conference*, Shavlik, J. ed., Morgan Kaufmann Publishers, San Francisco, CA, (1998)

[28] R. Fano., *Transmission of information*, MIT Press, (1961).

[29] T. Cover and J. Thomas, *Elements of Information Theory*, Wiley, (1991).

[30] R.O. Duda and P.E. Hart *Pattern Classification and Scene Analysis*, John Wiley & Sons, (1973).

[31] F. Attneave, "Information aspects of visual perception", *Psych. Rev.*, (61) 183-193, (1954).

[32] R. Blahut, *Principles and Practice of Information Theory,* Addison Wesley, (1987).

[33] D. E. Rumelhart, G. E. Hinton, and J. R. Williams, "Learning representations by back-propagating errors", *Nature* (London), 323, 533-536, (1986).

[34] H. Jeffreys, *Theory of probability*, Oxford, (1939).

[35] G. Deco, *An Information-Theoretic Approach to Neural Computing*, New York, Springer, (1996).

[36] J. W. Fisher, *Nonlinear Extensions to the Minimum Average Correlation Energy Filter* Ph.D dissertation, University of Florida, (1997)

[37] C. E. Shannon, "A mathematical theory of communication". *Bell Sys. Tech. J.* 27, 379-423, (1948).

[38] A. Renyi, "Some Fundamental Questions of Information Theory". *Selected Papers of Alfred Renyi*, 2, 526-552, Akademiai Kiado, Budapest, (1976).

[39] A. Renyi, A. "On Measures of Entropy and Information". *Selected Papers of Alfred Renyi,* 2, 565-580, Akademiai Kiado, Budapest, (1976).

[40] E. Parzen, "On the estimation of a probability density function and the mode", *Ann. Math. Stat.* 33, 1065-1076, (1962).

[41] S. Kullback, *Information Theory and Statistics*, Dover Publications, Inc., New York, (1968).

[42] D. X. Xu, *Energy, Entropy and Information Potential for Neural Computation*, Ph.D dissertation, University of Florida, (1998).

[43] E. Jaynes, "Information theory and statistical mechanics", *Phys. Rev.*, 106, 620-630, (1957).

[44] R. Fisher, *Statistical methods and Scientific Inference*, Hafner, New York, (1956).

[45] G. Strang and T. Nguyen. *Wavelets and Filter Banks.* Wellesley-Cambridge Press, Wellesley, MA, (1996).

[46] S. Y. Kung. *Digital Neural Networks*. PTR Prentice Hall, Englewood Cliffs, NJ, (1993).

[47] D. Xu, J. Fisher, and J. Principe, "Mutual information approach to pose estimation", *Algorithms for synthetic aperture radar imagery V*, E. Zelnio, Eds., Proceedings of the SPIE, vol 3370, 218-229. (1998).

[48] Q. Zhao, D.X. Xu, and J. Principe. "Pose estimation of SAR automatic target recognition." Proceedings of Image Understanding Workshop, Monterey, CA., 827-832, (1998).

[49] A.W. Learn, *Target pose estimation from radar data using adaptive networks*, Master thesis, Air Force Institute of Technology. (1999)

[50] J. Principe, Q. Zhao and D. Xu. "A novel ATR classifier exploiting pose information". In Proceedings of Image Understanding Workshop, 833-836, Monterey, CA., (1998).

[51] J. Principe, D. Xu, and J. Fisher, "Information Theoretic Learning", Chapter in *Unsupervised Adaptive Filtering*, Simon Haykin Editor, Wiley, (1999).

[52] Q. Zhao and J. Principe, "From hyperplanes to large margin classifiers: applications to SAR ATR", *Automatic Target Recognition IX*, F. Sadjadi Ed. Proceedings of the SPIE, 3718, 101-109. (1999)

[53] Q. Zhao and J. Principe, "Forming large margins with support vector machines for synthetic aperture radar automatic target recognition", submitted to *Optical Engineering*, 1999

[54] J. Principe, D. Xu., and Q. Zhao, "Learning from examples with information theoretic criteria", submitted to *VLSI Signal Processing Systems*, (1999).

[55] J. Principe, M. Kim, and J. Fisher, "Target detection in synthetic aperture radar (SAR) using artificial neural networks", *IEEE Trans. Image Proc*. (special issue on neural networks), 7(8), 1136-1149, (1998).

[56] M. Bryant and F. Garber. "SVM classifier applied to the MSTAR public data set", *Algorithms for Synthetic Aperture Radar Imagery VI*, E. Zelnio, Eds., Proceedings of the SPIE, 3721, 355-360. (1999)

Appendix

Information Potential for Discrete Samples

Let $a_i \in R^k$, $i = 1, ..., N$, be a set of samples from a random variable $Y \in R^k$ in k-dimensional space. One interesting question is what will be the entropy associated with this set of data points. One answer lies in the estimation of the data pdf by the Parzen window method using a Gaussian kernel [40]:

$$f_Y(y) = \frac{1}{N} \sum_{i=1}^{N} G(y - a_i, \sigma^2) \qquad (18)$$

where $G(\ ,\ )$ is the Gaussian kernel as above and σ^2 is the variance. When Shannon's entropy is used along with this pdf estimation, the measure becomes very complex to compute. Fortunately, Renyi's entropy with order 2 or quadratic entropy leads to a simpler form by using (10) and we obtain the entropy measure for a set of discrete data points $\{a_i\}$ as:

$$H(\{a_i\}) = H_{R2}(Y|\{a_i\}) = -\log\left(\int_{-\infty}^{+\infty} f_Y(y)^2 dy\right) = -\log V(\{a_i\})$$

$$\qquad (19)$$

$$V(\{a_i\}) = \frac{1}{N^2} \sum_{i=1}^{N} \sum_{j=1}^{N} \int_{-\infty}^{+\infty} G(y - a_i, \sigma^2) G(y - a_j, \sigma^2) dy = \frac{1}{N^2} \sum_{i=1}^{N} \sum_{j=1}^{N} G(a_i - a_j, 2\sigma^2)$$

Making the analogy between data points and "particles", $V(\{a_i\})$ can be regarded as an overall potential energy since $G(a_i - a_j, 2\sigma^2)$ can be taken as the potential energy of "particle" a_i in the potential field of "particle" a_j, or vice versa. *We will call this potential energy an information potential.* So, maximizing entropy in this case is equivalent to minimizing information potential.

Information Forces

Just like in mechanics, the derivative of the potential energy is a force, in this case an information driven force that moves the data samples in the space of the interactions. Therefore,

$$\frac{\partial}{\partial a_i} G(a_i - a_j, 2\sigma^2 I) = -G(a_i - a_j, 2\sigma^2 I)(a_i - a_j)/(2\sigma^2) \qquad (20)$$

can be regarded as the force F_{ij} that the sample a_j impinges upon a_i, and *will be called an information force (IF).* If we add all the contributions of the IF from the ensemble of samples on a_i we have the net effect of the information potential on sample a_i, i.e.

$$\nabla_i = \frac{\partial}{\partial a_i} V(y) = -\frac{1}{N^2\sigma^2} \sum_{j=1}^{N} G(a_i - a_j, 2\sigma^2 I)(a_i - a_j) = \frac{-1}{N^2\sigma^2} \sum_{j=1}^{N} V_{ij} d_{ij} \qquad (21)$$

"Force" Back-Propagation

The concept of IP creates a criterion, which is external to the mapping topology (as an MLP). The only missing step is to integrate the criterion with the adaptation of a parametric mapper as the MLP. Suppose the samples y are the outputs of the MLP. If we want to adapt the MLP such that the mapping maximizes the entropy at the output $H(y)$, the problem is to find the MLP parameters w_{ij} so that the IP $V(y)$ is minimized. In this case, the IPCs are not free but are a function of the MLP parameters. So, the information forces applied to each IPC by the information potential can be back-propagated to the parameters using the chain rule, i.e.

$$\frac{\partial}{\partial w}V(y) = \sum_{i=1}^{N}\left[\frac{\partial}{\partial a_i}V(y)\right]^T\frac{\partial a_i}{\partial w} = \sum_{i=1}^{N} F_i^T\frac{\partial}{\partial w}g(w, x_i) \tag{22}$$

where $a_i = (a_{i1}, ..., a_{iM})^T$ is the M-dimensional MLP output. Notice that from (22) the sensitivity of the output with respect to a MLP parameter $\dfrac{\partial y_i}{\partial w}$ is the "*transmission mechanism*" through which information forces are back-propagated to the parameter.

Cross-Information Potential for Discrete Samples

Now, suppose that we observe a set of data samples $\{a_{i1}, i= 1, ..., N\}$ for the variable Y_1, $\{a_{i2}, i= 1, ..., N\}$ for the variable Y_2. Let $a_i = (a_{i1}, a_{i2})^T$. Then $\{a_i, i= 1, ..., N\}$ are data samples for the joint variable $(Y_1, Y_2)^T$. Based on the Parzen window method, the joint pdf and marginal pdf can be estimated as:

$$f_{Y_1 Y_2}(y_1, y_2) = \frac{1}{N}\sum_{i=1}^{N} G(y_1 - a_{i1}, \sigma^2)G(y_2 - a_{i2}, \sigma^2)$$

$$f_{Y_1}(y_1) = \frac{1}{N}\sum_{i=1}^{N} G(y_1 - a_{i1}, \sigma^2) \tag{23}$$

$$f_{Y_2}(y_2) = \frac{1}{N}\sum_{i=1}^{N} G(y_2 - a_{i2}, \sigma^2)$$

Combining (20), (21) and using (11), we obtain the following expressions for the Quadratic Mutual Information based on a set of data samples:

$$V_{ED}(y) = V(\{a\}) - 2V_{nc}(\{a\}) + V^1(\{a_1\})V^2(\{a_2\})$$

$$V(\{a_i\}) = \frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N} G(a_i - a_j, 2\sigma^2) = \frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}\left(\prod_{l=1}^{2} G(a_{il} - a_{jl}, 2\sigma^2)\right)$$

$$V^l(a_j, \{a_{il}\}) = \frac{1}{N}\sum_{i=1}^{N} G(a_{jl} - a_{il}, 2\sigma^2), \quad l = 1, 2$$

$$V^l(\{a_{il}\}) = \frac{1}{N}\sum_{j=1}^{N} V^l(a_j, \{a_{il}\}), \qquad l = 1, 2 \tag{24}$$

$$V_{nc}(\{a_i\}) = \frac{1}{N}\sum_{j=1}^{N}\left(\prod_{l=1}^{2} V^l(a_j, \{a_{il}\})\right)$$

In order to interpret these expressions in terms of information potentials we have to make some further definitions: We will use the term marginal when the information potential is calculated in a subspace, and partial when only some of the data points are used. With this in mind $V(\{a_i\})$ is the overall information potential in the joint space, $V_l(a_j, \{a_{il}\})$ is the partial marginal information potential because it is the potential of the point a_j in its corresponding marginal information potential field (indexed by l). $V_l(\{a_{il}\})$ is the marginal information potential because it averages all the partial marginal information potentials for one index l, and $V_{nc}(\{a_i\})$ is the un-normalized cross-information potential because it measures the interactions between the partial marginal information potentials.

"Forces" in the Cross-Information Potential

The cross-information potential is more complex than the information potential. Three different potentials contribute to the cross-information potential. So, the force applied to each data point a_p comes from three independent sources. A force in the joint space can decomposed into marginal components. The marginal force of q (marginal space indexed by q) that the data point a_p received from three sources can be calculated according to the following formula:

$$\frac{\partial}{\partial a_{pq}} V(\{a_i\}) = \frac{1}{N^2} \sum_{i=1}^{N} \left[\prod_{l=1}^{k} G(a_{iq} - a_{pq}, 2\sigma^2) \right] \frac{a_{iq} - a_{pq}}{\sigma^2}$$

$$\frac{\partial}{\partial a_{pq}} V_q(\{a_{iq}\}) = \frac{1}{N^2} \sum_{i=1}^{N} G(a_{iq} - a_{pq}, 2\sigma^2) \frac{a_{iq} - a_{pq}}{\sigma^2} \qquad (25)$$

$$\frac{\partial}{\partial a_{pq}} V_{nc}(\{a_i\}) = \frac{1}{N^2} \sum_{i=1}^{N} \frac{1}{2} \left(\prod_{l \neq a} V_l(a_j, \{a_{il}\}) + \prod_{l \neq q} V_l(a_p, \{a_{il}\}) \right) G(a_{jq} - a_{pq}) \frac{a_{jq} - a_p}{\sigma^2}$$

It is also not difficult to obtain the formula for the calculation of the information force produced by the CIP field in the case of the Euclidean difference measure

$$c_{ij}^k = V_{ij}^k - V_i^k - V_j^k + V^k, \quad k = 1, 2$$

$$F_i^l = \frac{\partial V_{ED}}{\partial y_i^l} = \frac{-1}{N^2 \sigma^2} \sum_{j=1}^{N} c_{ij}^k V_{ij}^l d_{ij}^l \qquad (26)$$

$$i = 1, ..., N, \quad l \neq k, \quad l = 1, 2$$